Locality Based Analysis of Network Flows

SEI/CERT

21 July 2004

John McHugh,
Carrie Gates, Damon Becknel

© 2004 by Carnegie Mellon University

Why Locality

- Locality is an entropy based characterization that allows prediction of future behavior based on past observations.
  - It captures the degree to which the behavior of a system is regular in some sense
  - It appears to be scale free, appearing in internet, subnet, and node scale behaviors.
  - It promotes clustering allowing the use of sets and multisets to abstract group behaviors.

© 2004 by Carnegie Mellon University

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
21 JUL 2004		**00-00-2004 to 00-00-2004**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Locality Based Analysis of Network Flows	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Carnegie Mellon University,Software Engineering Institute,Pittsburgh,PA,15213**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
presented at FloCon 2004, Crystal City, VA, July 2004.

14. ABSTRACT

15. SUBJECT TERMS

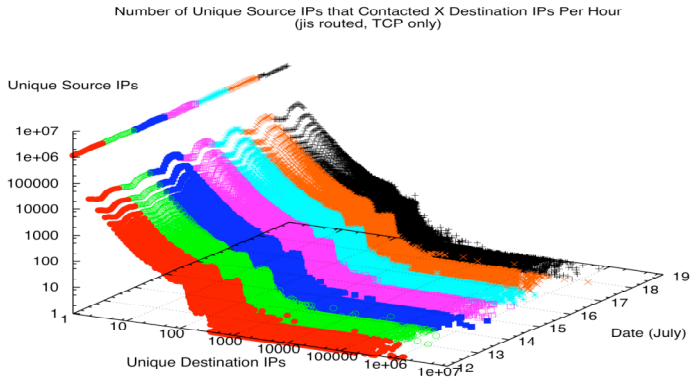
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**13**	

Eye Candy vs. Insight

- Locality often manifests as patterns in some space.
  - If we select the appropriate dimensions, we may achieve either understanding or puzzlement.
  - The next three pictures show persistent structure where none might be expected.
  - This can be viewed as a summary of a time series of connection matrices.
  - Graphics by Carrie Gates

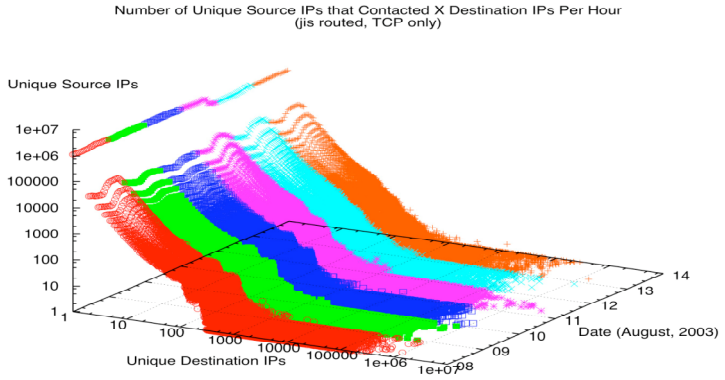
© 2004 by Carnegie Mellon University

First you see it ...



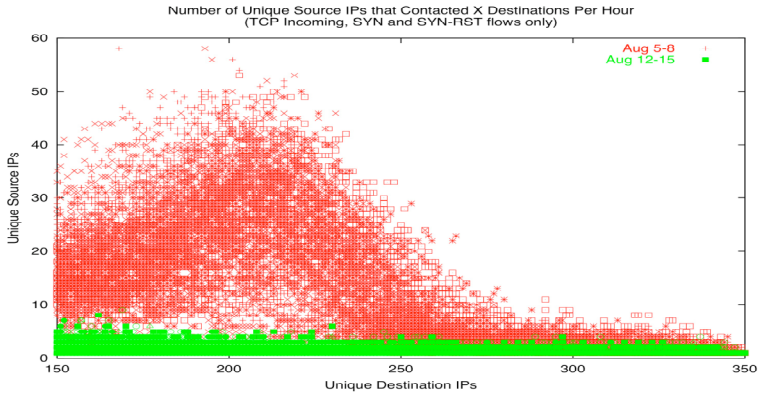
© 2004 by Carnegie Mellon University

Then it goes away ...



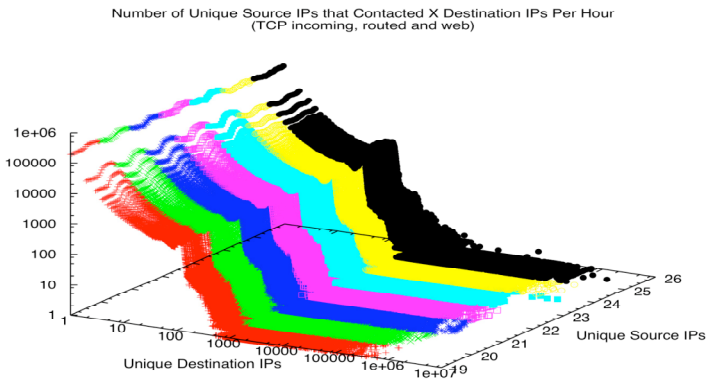
© 2004 by Carnegie Mellon University

(rather abruptly)



© 2004 by Carnegie Mellon University

Only to return (months later).



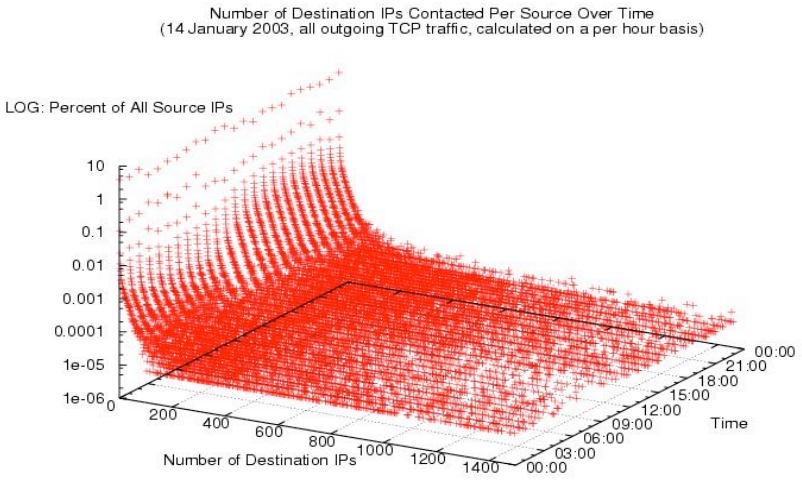
© 2004 by Carnegie Mellon University

Williamson's Locality

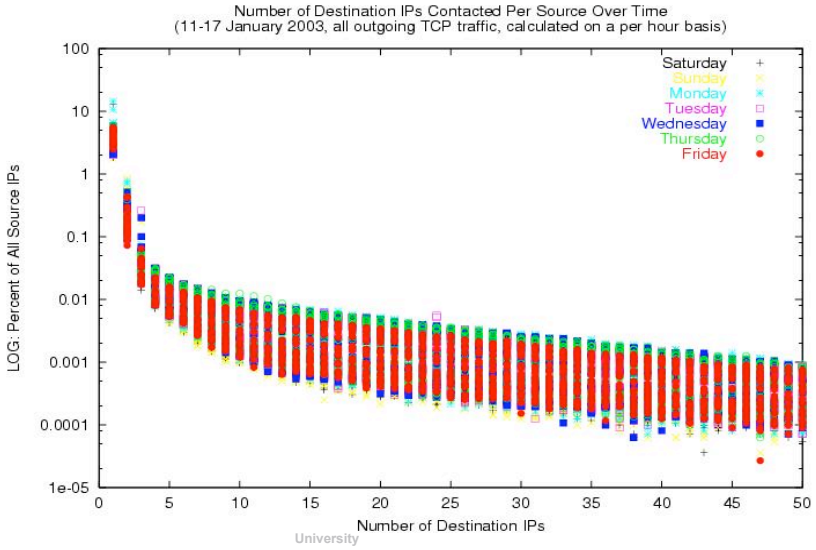
- Matt Williamson, late of HP Bristol, noted address locality in a 2002 ACSAC paper.
  - For browsing, last 10 IPs visited constitute an effective working set.
  - Working set violations relatively rare, bursts rarer yet.
    - Delay on violation is effective "soft" mitigator
- What is the locality of trans border data?

© 2004 by Carnegie Mellon University

Detail of Inside to Outside Day



Weekly In/Out Locality Range

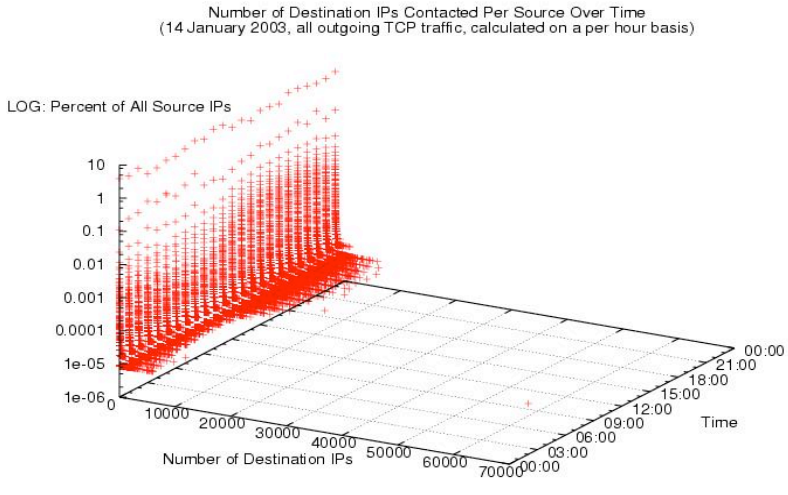


Williamson Confirmed (mostly)

- With the caveat that we are not seeing internal connections, the vast majority of the flows arguably follow Williamson's working set model.
- As usual, there are outliers ...

© 2004 by Carnegie Mellon University

One Day of Inside to Outside



University

Noise localities

- We have been characterizing modest subnets in support of the traffic generation that will be used in the DARPA DQ system evaluations.
  - Attempting to avoid mistakes of DARPA IDS evaluation.
  - Striving for a realistic noise environment, among other things.

© 2004 by Carnegie Mellon University

Carnegie Mellon University
Software Engineering Institute

Crud and Noise

• In January, we observed a /16 for a week, and the whole customer net for a minute

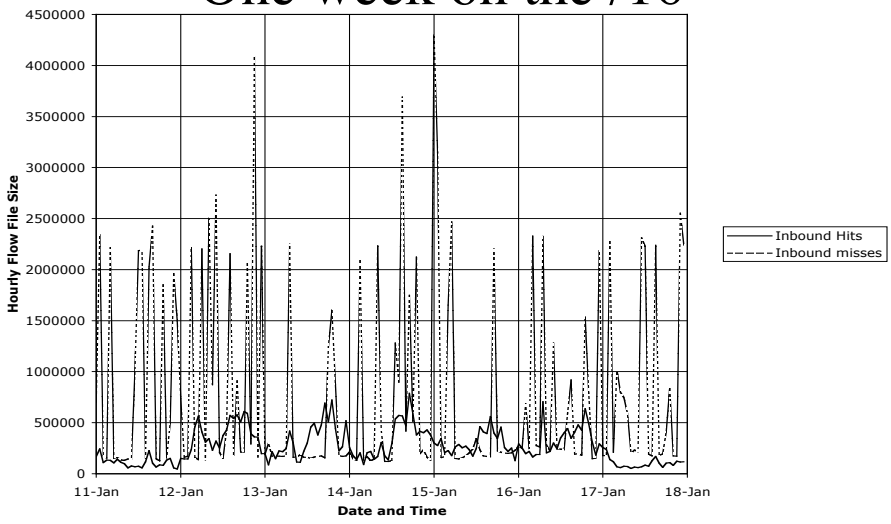
•For the /16

- MMM.NNN.24.x - 66 hosts MMM.NNN.25.x - 60 hosts
- MMM.NNN.26.x - 46 hosts MMM.NNN.27.x - 49 hosts
- MMM.NNN.28.x - 57 hosts MMM.NNN.29.x - 7 hosts
- MMM.NNN.30.x - 70 hosts MMM.NNN.31.x - 67 hosts
- MMM.NNN.32.x - 54 hosts MMM.NNN.33.x - 62 hosts
- MMM.NNN.34.x - 50 hosts MMM.NNN.35.x - 4 hosts
- MMM.NNN.120.x - 2 hosts MMM.NNN.127.x - 1 host
- MMM.NNN.140.x - 1 host MMM.NNN.251.x - 4 hosts
- Total 600 hosts in 16 /24s

© 2004 by Carnegie Mellon University

Carnegie Mellon University
Software Engineering Institute

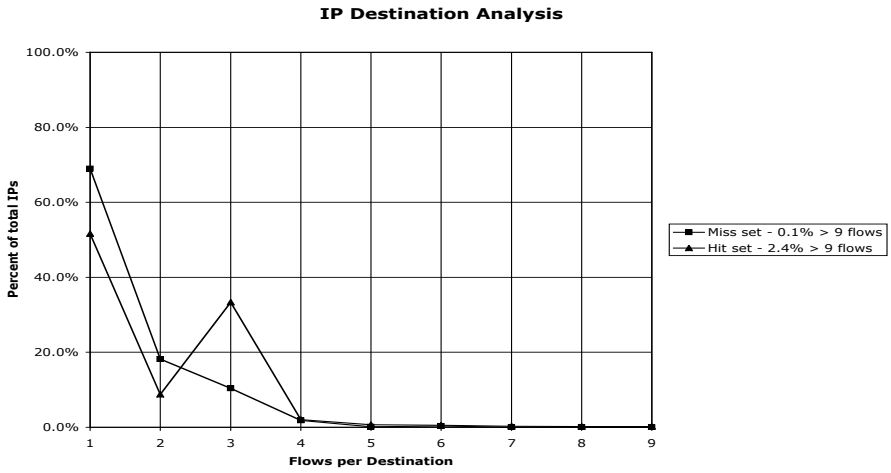
One week on the /16



© 2004 by Carnegie Mellon University

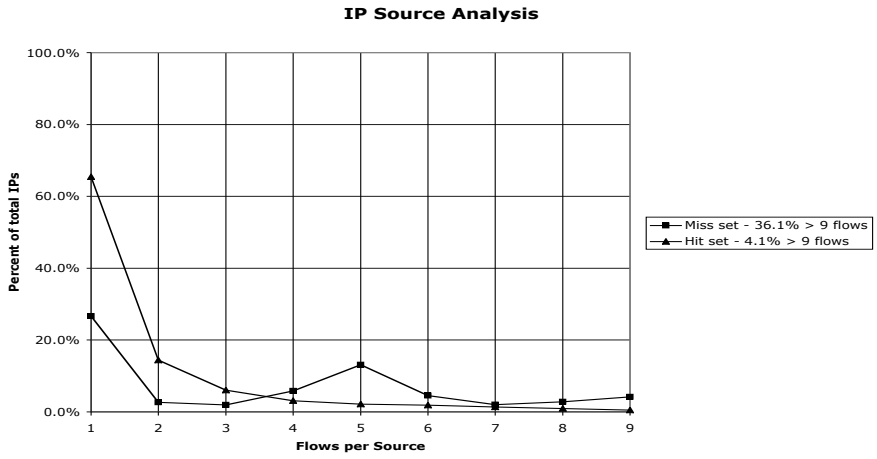
Carnegie Mellon University
Software Engineering Institute

1 Min sample - destinations



© 2004 by Carnegie Mellon University

1 Min Sample - sources



© 2004 by Carnegie Mellon University

top 5 in 1 min sample

- Created a “bag” for source and destination addresses in the 1 minute sample. The annotated top 5 are:
- (39) lip $ readbag --count --print jcm-tcp-s-10+.bagl sort -r -n | head
  - 12994 AAA.BBB.068.218 - scan 4899 (Radmin)
  - 6598 CCC.DDD.209.215 - scan 7100 (X-Font)
  - 5944 EEE.FFF.125.117 - scan 20168 (Lovegate)
  - 5465 GGG.HHH.114.052 - ditto
  - 5303 III.JJJ.164.126 - scan 3127 (My doom)

© 2004 by Carnegie Mellon University

Bottom of bag in 1 min sample

- 3335 external hosts sent exactly one TCP flow
  - SYN probes for port 8866 449 times
    - W32.Beagle.B@mm is a mass-mailing worm-back door on TCP port 8866.
  - SYN probes for port 25 are seen 271 times.
  - Most remainder are SYNs to a variety of ports, mostly with high port numbers.
  - There are a number of ACK/RST packets which are probably associated with responses to spoofed DDoS attacks.

© 2004 by Carnegie Mellon University

Carnegie Mellon University
Software Engineering Institute

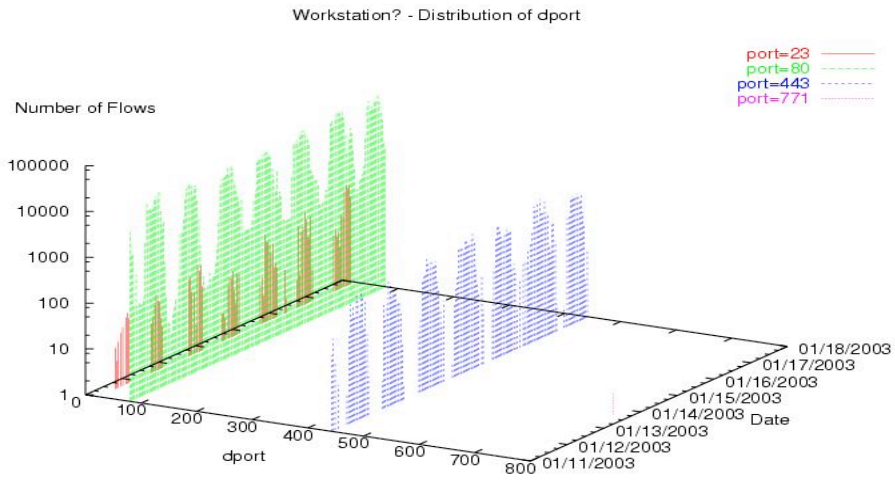
Individual host profiles

- These were done by Capt. Damon Becknel, USA.
  - He was looking for ways of characterizing the role of a node based on it's activity patterns
  - As usual, surprising results are sometimes observed.

© 2004 by Carnegie Mellon University

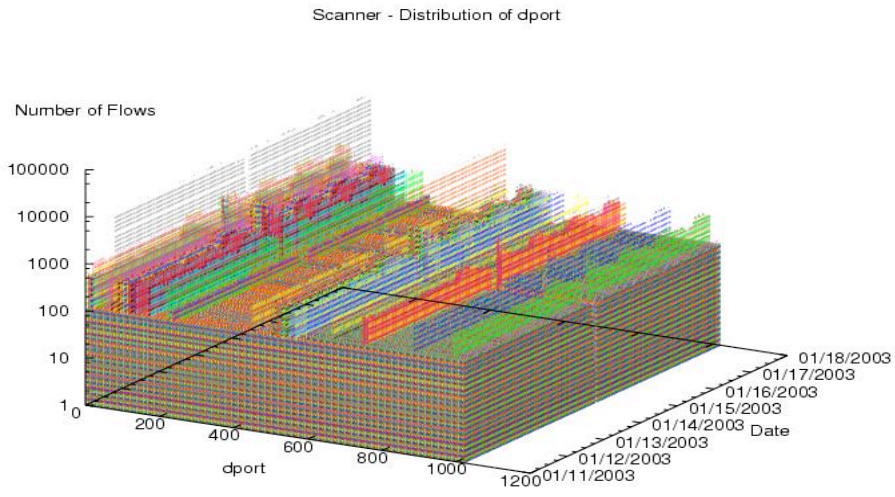
Carnegie Mellon University
Software Engineering Institute

Workstation?



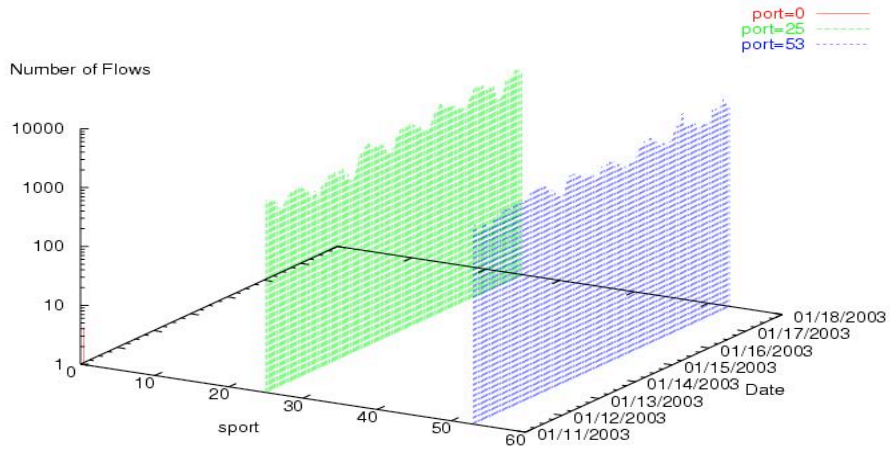
Carnegie Mellon University
Software Engineering Institute

Scanner



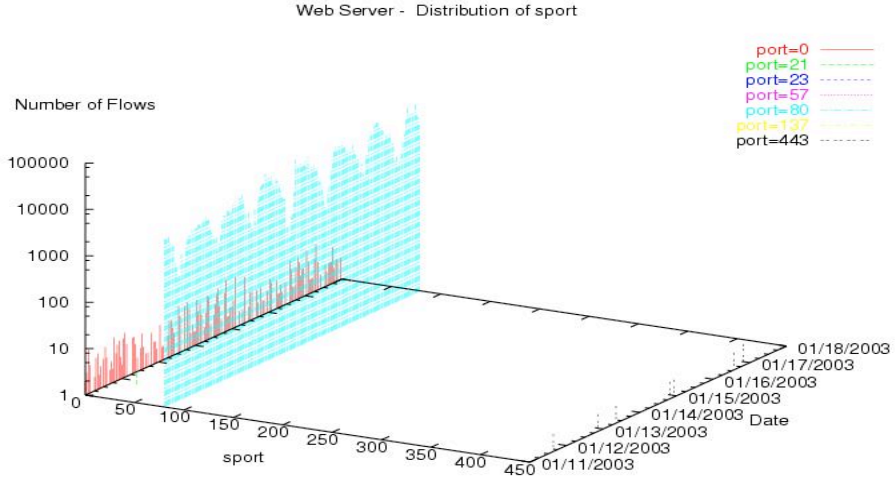
Carnegie Mellon University
Software Engineering Institute

Mail Server?



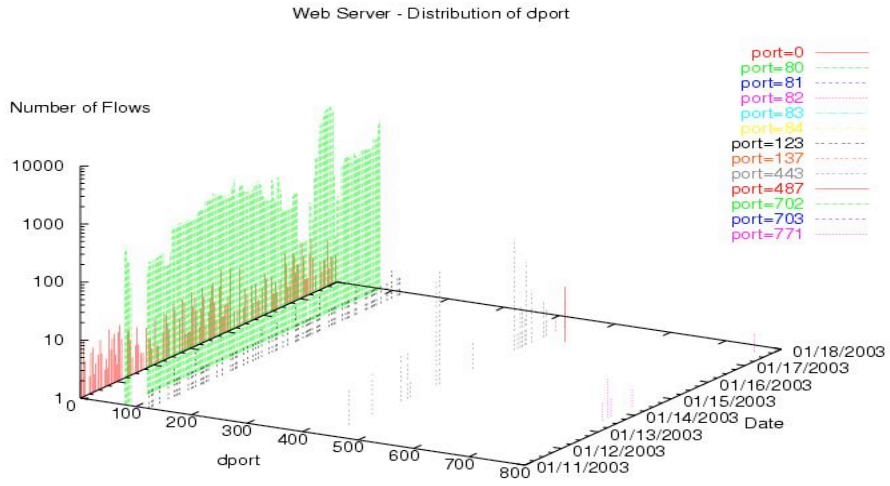
Carnegie Mellon University
Software Engineering Institute

Web Server



Carnegie Mellon University
Software Engineering Institute

Web Server



Carnegie Mellon University
Software Engineering Institute

Summary

- We have provided some examples of locality on a variety of scales for a variety of representations.
- It is our hope that the general notions of locality, and clustering will provide a basis for reducing the complexity of analysis.

© 2004 by Carnegie Mellon University